Our Ref.: 042390.P16531

APPLICATION FOR UNITED STATES LETTERS PATENT

FOR

Method and System for a Network Node for Attachment to Switch Fabrics

Inventors: **Neal Oliver**
David Gish
Gerald Lebizay
Henry Mitchel
Brian Peebles
Alan Stone

Prepared by:

BLAKELY SOKOLOFF TAYLOR & ZAFMAN, LLP
12400 Wilshire Boulevard, 7th Floor
Los Angeles, California 90025
(503) 684-6200

Express Mail No.: EV325525651US

Method and System for a Network Node for Attachment to Switch Fabrics

BACKGROUND

1. Technical Field

[0001] Embodiments of the invention relate to the field of networking, and more specifically to a network node for attachment to switch fabrics.

2. Background Information and Description of Related Art

[0002] Many modern communications applications have demanding resource requirements that are difficult to scale. An effective way to scale such applications is to distribute individual algorithms onto different processing elements and interconnect those processing elements on a switch fabric. This arrangement may support the high-bandwidth data flows that must exist between the algorithms.

[0003] A high performance switch fabric arrangement may be achieved by use of various proprietary multi-stage switching technologies, but the performance is achieved at a high cost in dollars and in features, such as scalability. In order to achieve high performance at a reasonable cost, current approaches include the use of network technologies such as Ethernet, InfiniBand, PCI-Express/Advanced Switching, and Rapid IO. However, these network technologies are immature and still evolving and are limited in some features critical to network performance, such as classes of service and congestion control.

BRIEF DESCRIPTION OF DRAWINGS

[0004] The invention may best be understood by referring to the following description and accompanying drawings that are used to illustrate embodiments of the invention. In the drawings:

[0005] **FIG. 1** is a block diagram illustrating one generalized embodiment of a system incorporating the invention.

[0006] **FIG. 2** is a block diagram illustrating one generalized embodiment of a system incorporating the invention in greater detail.

[0007] **FIG. 3** illustrates a hardware architecture of a network node according to one embodiment of the invention.

[0008] **FIG. 4a** illustrates an interconnection of nodes in a multishelf configuration using an external switch according to one embodiment of the invention.

[0009] **FIG. 4b** illustrates an interconnection of nodes in a multishelf configuration using a mesh according to one embodiment of the invention.

[0010] **FIG. 5** is a flow diagram illustrating a method according to an embodiment of the invention.

DETAILED DESCRIPTION

[0011] Embodiments of a system and method for a network node for attachment to switch fabrics are described. In the following description, numerous specific details are set forth. However, it is understood that embodiments of the invention may be practiced without these specific details. In other instances, well-known circuits, structures and techniques have not been shown in detail in order not to obscure the understanding of this description.

[0012] Reference throughout this specification to “one embodiment” or “an embodiment” means that a particular feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment of the invention. Thus, the appearances of the phrases “in one embodiment” or “in an embodiment” in various places throughout this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures, or characteristics may be combined in any suitable manner in one or more embodiments.

[0013] Referring to Fig. 1, a block diagram illustrates a network node 100 according to one embodiment of the invention. Those of ordinary skill in the art will appreciate that the network node 100 may include more components than those shown in Fig. 1. However, it is not necessary that all of these generally conventional components be shown in order to disclose an illustrative embodiment for practicing the invention.

[0014] Network node 100 includes a switch 104 to couple to a switch fabric 102 and a plurality of subsystems, such as 106, 108, and 110. The subsystem 106 is a subsystem at which external traffic, such as ATM virtual circuits, SONET, and Ethernet, enters and exits the network node 100. The subsystem 108 labels each received external packet to identify an associated flow, and classifies each external packet into one of a plurality of queues based on the packet's destination, priority, and path through the switch fabric 102. The subsystem 110 receives labeled and classified packets, maps the packets into the appropriate queue, schedules the packets from each queue for transmission, and encapsulates the packets to form frames of uniform size before transmitting the packets to the switch fabric 102 through switch 104.

[0015] In one embodiment, the network node 100 also includes one or more adjunct subsystems that perform various high-touch processing functions, such as deep packet inspection and signal processing. A packet may be routed to an internal or external adjunct subsystem for processing. An adjunct process may be a thread of a network processor core, a thread of a network processor microengine, or a thread of an adjunct processor, such as a digital signal processor (DSP). The adjunct process may be on a local node or an external node.

[0016] Although the exemplary network node 100 is shown in Fig. 1 and Fig. 2 as including a switch 104 to connect the subsystems and the switch fabric, in one embodiment, the switch 104 could be split into two switches. One of the two switches would be a local switch that connects the various subsystems of the network node. The other of the two switches would be a fabric switch that connects one or more subsystems to the switch fabric.

[0017] Fig. 2 illustrates the subsystems of network node 100 in greater detail according to one embodiment of the invention. As shown, subsystem 106 includes an input Media Access Control (MAC) 202 and an output MAC 204 to interface with external networks, such as ATM virtual circuits, SONET, and Ethernet. The subsystem 106 converts incoming data to packet streams, and formats and frames outbound packet streams for the network interface.

[0018] The subsystem 108 includes an input MAC 212, an output MAC 206, a classification function 208, and a decapsulation function 210. If an encapsulated frame is received at subsystem 108 from the switch fabric, it is sent to the decapsulation function 210, where the frame is decapsulated into the original packets. If an external packet is received at subsystem 108, then the external packet is sent to the classification function 208 to be labeled and classified.

[0019] The classification function 208 examines each external packet and gathers information about the packet for classification. The classification function 208 may examine a packet's source address and destination address, protocols associated with the packet (such as UDP, TCP, RTP, HTML, HTTP), and/or ports associated with the packet. From this information, the classification function 208 determines a particular flow associated with the packet and labels the packet with a flow identifier (ID) to identify the associated flow. The packet may then be classified into one of a plurality of traffic classes, such as voice, email, or video traffic. A path to be taken by the packet through the switch fabric is determined. Load balancing is considered when determining the paths packets will take through the switch fabric. Load balancing refers to selecting different paths for different flows to balance the load on the paths and to minimize the damage that could be done to throughput by a partial network failure.

[0020] Each packet is classified into one of a plurality of queues based on the packet's destination, path through the switch fabric, and priority. The packets in a queue have a common destination, path through the switch fabric, and priority. Each packet may be labeled with a queue ID to identify the queue to which the packet has been classified. In one embodiment, packets may be further edited by removing headers and layer encapsulations that are not needed during transmission through the system. After a packet is labeled and classified, it is sent back to switch 104 to be routed to subsystem 110.

[0021] The subsystem 110 includes an output MAC 214, an input MAC 222, a mapping element 216, a scheduler 218, and an encapsulation element 220. The mapping element 216 examines each packet and determines which one of a plurality of queues the packet belongs based on packet's label identifiers. The packet is then queued into the appropriate queue to await transmission to a next destination through the switch fabric. The scheduler 218 schedules the packets in the queues for transmission. The scheduler 218 uses various information to schedule packets from the queues. This information may include occupancy statistics, flowspec information configured via an administrative interface, and feedback from switch function. Various algorithms may be used for the scheduling, such as Longest Delay First, Stepwise QoS Scheduler (SQS), Simple Round Robin, and Weighted Round Robin.

[0022] After the packets have been dequeued and scheduled for transmission, the scheduler 218 sends the packets to the encapsulation element 220. The encapsulation element 220 transforms the scheduled packets into uniform size frames by aggregating small packets and segmenting large packets. The size of the frame may be determined by the Message Transfer Unit (MTU) of the switch fabric technology used in the system. Small packets may be merged together using multiplexing, while large packets may be divided up using segmentation and reassembly (SAR). The encapsulation also includes conveyance headers that contain information required to decode the frame back into the original packets. The headers may also include a sequence number of packets within a frame to aid in error detection and a color field to indicate whether a flow conforms with its flowspec.

[0023] The encapsulated frames are sent to input MAC 222, which translates each frame into a format consistent with the switch fabric technology, and then sends each frame to a switch fabric port consistent with the path selected for the frame. Different switch fabric technologies and implementations may be used in the system, including Ethernet, PCI-Express/Advanced Switching, and InfiniBand technologies.

[0024] The following is an example of a path through the network node 100 taken by an external packet received at subsystem 106. The external packet is received from an external network at the input MAC 202 in subsystem 106. The packet is sent to switch 104, which forwards the packet to subsystem 108 for classification. The packet arrives at MAC 206 in subsystem 108, which forwards the packet to the classification function 208. The classification function 208 examines the packet, determines a flow associated with the packet, labels the packet with a flow ID, determines a path to be taken by the packet through the switch fabric, and classifies the packet into one of a plurality of queues. The labeled and classified packet is then sent to MAC 212, which forwards the packet back to switch 104. The switch 104 sends the packet to subsystem 110. The packet arrives at MAC 214 in subsystem 110, which forwards the packet to the mapping element 216. The mapping element 216 examines the packet's label identifiers and determines which one of a plurality of queues the packet belongs. The packet is then queued into the appropriate queue to await transmission to a next destination through the switch fabric. The scheduler 218 schedules the packet in the queue for transmission. When the packet is scheduled for transmission and dequeued, the packet is encapsulated by the encapsulation function 220 into a uniform size frame by aggregating the packet with other packets if the packet is small or segmenting the packet if the packet is large. The frame is then sent to the MAC 222, which translates the frame into a format consistent with the switch fabric technology, and then sends the frame to a switch fabric port consistent with the path selected for the frame. The packet may then arrive at another network node similar to the one from which it was transmitted.

[0025] The following is an example of a path through the network node 100 taken by a frame received from the switch fabric 102. The frame is received at the switch 104. The frame is sent to MAC 206 in subsystem 108, which forwards the packet to the decapsulation function 210. The decapsulation function 210 decapsulates the frame into the original one or more packets. The packets are then sent back to the switch 104 to be forwarded locally or externally. For example, the switch may send the packet to an adjunct subsystem for high touch processing or to subsystem 106 to be transmitted to an external network.

[0026] Fig. 3 illustrates a hardware representation of a network node 300 according to one embodiment of the invention. The center of the node is a switch 302 that connects the node to the rest of the network via the switch fabric 304, and to various processing elements located on a baseboard and mezzanine boards. A PCI-Express/Advanced Switching Node is used in this exemplary implementation. However, other network technologies, such as Ethernet, and InfiniBand technologies may be used in the network node in other embodiments. In one embodiment, subsystem 106 and an external adjunct subsystem may be located on mezzanine boards while subsystems 108 and 110 and an internal adjunct subsystem are located on the baseboard.

[0027] Fig. 4a illustrates how a network node may be interconnected in a scalable system to additional switching nodes in a network according to one embodiment of the invention. Fig. 4b illustrates how a network node may be interconnected in a scalable system with individual boards connected directly in a mesh according to one embodiment of the invention. Every board need not be connected vertically, and other mesh arrangements may be used to connect the boards in other embodiments of the invention.

[0028] Fig. 5 illustrates a method according to one embodiment of the invention. At 500, each received network packet is labeled with information identifying an associated flow and a queue in which the packet will await transmission. In one embodiment, the flow associated with the packet is determined based on the source address, destination address, ports, and/or protocols associated with the packet. In one embodiment, the traffic class to which the packet belongs is determined. In one embodiment, header information and/or layer encapsulations are removed from the packet. At 502, each packet is placed into one of a plurality of queues to await transmission based on the packet's label identifiers. At 504, the packets in the queues are scheduled for transmission. At 506, the scheduled packets are encapsulated to form frames of uniform size by grouping small packets and segmenting large packets. In one embodiment, headers are added that contain information to decode each frame back into the original packets. In one embodiment, small packets are merged to form one frame using multiplexing. In one embodiment, large packets are divided up and placed into multiple frames using segmentation and reassembly. At 508, the uniform frames are transmitted through a switch fabric to a next destination. In one embodiment, when the frames are received at another network node, each frame is decapsulated into the original packets.

[0029] While the invention has been described in terms of several embodiments, those of ordinary skill in the art will recognize that the invention is not limited to the embodiments described, but can be practiced with modification and alteration within the spirit and scope of the appended claims. The description is thus to be regarded as illustrative instead of limiting.